Recall from Lecture 4, where the RUM model is introduced, we model the random utility by decomposing it into deterministic and random components, i.e.

$$U_{ij} = V_{ij} + \varepsilon_{ij},$$

where

$$U_{ij} : \text{random utility from item } j \text{ of customer } i$$
$$V_{ij} : \text{deterministic utility from item } j \text{ of customer } i$$
$$\varepsilon_{ij} : \text{error term from item } j \text{ of customer } i$$

The error term ε_{ij} helps us to capture the differences across people when we try to model the whole population, or even if only an individual is considered, it describes some uncertainty from those unobserved features.

1 Deterministic Component

Let's first focus on the deterministic component V_{ij}. As discussed in last lecture, it can be expressed in terms of observable features as follows,

$$V_{ij} = \alpha_j + \underline{\beta}^\top \underline{z}_{ij} + \underline{\gamma}_j^\top \underline{s}_i + \underline{\delta}_j^\top \underline{w}_{ij},$$

where

$$\alpha_j : \text{intercept term/alternative specific constant(ASC).}$$
$$\underline{\beta} : \text{generic coefficients}$$
$$\underline{z}_{ij} : \text{features that describe product } j \text{ at occurence } i$$
$$\underline{\gamma}_j : \text{alternative-specific coefficients for product } j$$
$$\underline{s}_i : \text{attributes that describe the occurence } i$$
$$\underline{\delta}_j : \text{alternative-specific coefficients for product } j$$
$$\underline{w}_{ij} : \text{features that describe product } j \text{ at occurence } i$$

Notice that we **do not** model the coefficients at individual level at this point. This is because it is usually hard to collect enough information for each individual to introduce the hierarchical model, say, Mixture Logit Model. Before moving on to interpret these coefficients, let's first make some remarks on this model.

1. Does it make sense to introduce generic coefficients with individual-specific attributes, i.e. introducing an additional term $\underline{\eta}^\top \underline{r}_i$ as follows,

$$V_{ij} = \alpha_j + \underline{\beta}^\top \underline{z}_{ij} + \underline{\gamma}_j^\top \underline{s}_i + \underline{\delta}_j^\top \underline{w}_{ij} + \underline{\eta}^\top \underline{r}_i?$$

For RUM model, shifting all the utilities by the same amount **does not** affect the choice behavior. In other words, only differences in utility matter as far as choice behavior is concerned. Notice that the difference is always

$$V_{ij} - V_{ij'} = (\alpha_j - \alpha_{j'}) + \underline{\beta}^\top (\underline{z}_{ij} - \underline{z}_{ij'}) + (\underline{\gamma}_j - \underline{\gamma}_{j'})^\top \underline{s}_i + (\underline{\delta}_j^\top \underline{w}_{ij} - \underline{\delta}_{j'}^\top \underline{w}_{ij'}),$$

regardless of the introduction of new term. Therefore, adding generic coefficients with individual-specific attributes does not result in any change in the choice behavior.

Example 1. Suppose there are two customers making choices among two products, and we observe their choice behaviors, $p_1 = \mathbb{P}_1(1|\{1,2\})$ and $p_2 = \mathbb{P}_2(1|\{1,2\})$ respectively. How would choice behavior of customer 1 change, if income increased by \$1?

To make life easier, we assume the deterministic component only depends on the income of customer i, r_i, i.e.

$$V_{11} = \alpha_1 + \eta r_1, \qquad\qquad V_{21} = \alpha_1 + \eta r_2,$$
$$V_{12} = \alpha_2 + \eta r_1, \qquad\qquad V_{22} = \alpha_2 + \eta r_2.$$

If their choice behaviors follow MNL model, then

$$p_1 = p_2 = \frac{e^{\alpha_1 - \alpha_2}}{1 + e^{\alpha_1 - \alpha_2}},$$

and we can directly estimate $\delta = \alpha_1 - \alpha_2$ from p_1, p_2.

However, if given $p_1 > p_2$, then MNL model can no longer describe their choice behaviors simultaneously. Assume it still holds for each individual, and we choose item 2 as the **reference item** by setting $V_{12} = V_{22} = \alpha_2$, then

$$p_1 = \frac{e^{\alpha_1 - \alpha_2 + \eta r_1}}{1 + e^{\alpha_1 - \alpha_2 + \eta r_1}} \Rightarrow \delta + \eta r_1 = \log(\frac{p_1}{1 - p_1}).$$

Similarly, we also have

$$\delta + \eta r_2 = \log(\frac{p_2}{1 - p_2}).$$

Solve for δ, η using above equations, then we can compute the probability of customer 1 purchasing item 1 with income r_1' as

$$\frac{e^{\delta + \eta r_1'}}{1 + e^{\delta + \eta r_1'}}.$$

2. The same logic applies to ASCs. If we include one in every item, then we cannot identify them. We can only identify their differences. Because of this, it is common to introduce ASCs and any generic coefficients with individual specific attributes in all but one **reference item**. In some applications, there may be a natural reference item, which is the no-purchase option.

3. A model with only **alternative-specific** attributes/variables is called **conditional logit model**. A model with only **individual-specific** attributes/variables is called **multinomial logit model**.

2 Interpretation of Coefficients

Unlike in linear regression, the coefficients of RUM model are not directly interpretable, because cardinal value of utilities are not directly observable. We will interpret the coefficients through the impact on choice probabilities.

I. **Marginal effect interpretation**

1. **Individual/Choice-specific variables($\underline{s}_i$):**

   (a) **Marginal effect on odds ratio:**

   Suppose γ_j is the coefficient of item j corresponding to an individual-specific variable s_i, say, income of customer i. Then e^{γ_j} captures the item's marginal impact on odds ratio of customer's choice behavior. More specifically, consider the ratio of
   $$\frac{\mathbb{P}(j \text{ is chosen})}{\mathbb{P}(n \text{ is chosen})} = e^{V_{ij}-V_{in}}.$$
   We set item n as the reference alternative, i.e. set $\gamma_n = 0$. If keep everything else fixed, but increase s_i by 1 unit, then the odds ratio would get multiplied by e^{γ_j}. If $\gamma_j > 0$, we can conclude the customer is more likely to buy the product.

   (b) **Marginal effect on purchase probability itself:**

   It can be shown that under MNL model,
   $$\frac{\partial \mathbb{P}_i(j)}{\partial s_i} = \mathbb{P}_i(j)\left[\gamma_j - \sum_k \mathbb{P}_i(k)\gamma_k\right].$$

   We assume everything is on offer and $\mathbb{P}_i(j)$ refers to the probability that customer i buys product j. The effect of marginal increase on the choice probability depends on whether γ_j is bigger(or smaller) relative to the average of other coefficients, i.e., $\sum_k \mathbb{P}_i(k)\gamma_k$.

2. **Alternate-specific variables($\underline{z}_{ij}$):**

   (a) **Marginal effect on self-elasticity:**
   $$\frac{\partial \mathbb{P}_i(j)}{\partial z_{ij}} = \begin{cases} \beta \mathbb{P}_i(j)[1 - \mathbb{P}_i(j)], & \text{if the corresponding coefficient is generic} \\ \delta_j \mathbb{P}_i(j)[1 - \mathbb{P}_i(j)], & \text{otherwise} \end{cases}.$$

   The sign of coefficient determines whether the probability goes up or down.

   (b) **Marginal effect on cross-elasticity:**
   $$\frac{\partial \mathbb{P}_i(j)}{\partial z_{ik}} = \begin{cases} -\beta \mathbb{P}_i(j)\mathbb{P}_i(k), & \text{if the corresponding coefficient is generic} \\ -\delta_k \mathbb{P}_i(j)\mathbb{P}_i(k), & \text{otherwise} \end{cases}.$$

   (c) **Marginal effect on percentage change:**

   $$\frac{\partial \log \mathbb{P}_i(j)}{\partial z_{ik}} = \begin{cases} -\beta \mathbb{P}_i(k), & \text{if the corresponding coefficient is generic} \\ -\delta_k \mathbb{P}_i(k), & \text{otherwise} \end{cases}.$$

   That is, the percentage change in choice probability **does not** depend on product j. This is another implication of IIA property.

3

II. **Marginal rates of substitution**

The coefficients can also be used to explain the marginal rates of substitution among those attributes. Take a specific example as follows.

Example 2. Assume $V = \alpha + \beta_1 z_1 + \beta_2 z_2$, say, z_1 is travel time in hours, z_2 is cost in dollars. Then β_1/β_2 gives us the dollar value put on an hour of travel time. To be more precise, suppose we write

$$V = \alpha + \beta_{\text{time}} \times \text{time} + \beta_{\text{cost}} \times \text{cost}.$$

We expect $\beta_{\text{time}} < 0$, $\beta_{\text{cost}} < 0$. If we decrease travel time by 1 hour, the utility of customer increases by $|\beta_{\text{time}}|$, when all else is held constant. We can increase the cost by $|\beta_{\text{time}}|/|\beta_{\text{cost}}|$ while still not decreasing the utility. Therefore, the user is willing to pay an additional $|\beta_{\text{time}}|/|\beta_{\text{cost}}|$ for an hour reduction in travel time.

3 Error Term

As of now, we all concentrate on the deterministic component V_{ij} in the model. Let's turn to discussing about the error term ε_{ij} a little bit. The MNL model assumes that the error terms are independent across the items and individuals. The problem with this assumption is that it requires **all** the unobservables to be independent across both items and individuals. However, it does not always hold in rea

Example 3. We have 4 options, car, carpool, bus and train. We observe travel time(t) and travel cost(c) for each mode of transportation and use the following model,

$$V_{\text{car}} = \beta_t \times t_{\text{car}} + \beta_c \times c_{\text{car}} + \varepsilon_{\text{car}}$$
$$V_{\text{carpool}} = \beta_t \times t_{\text{carpool}} + \beta_c \times c_{\text{carpool}} + \varepsilon_{\text{carpool}}$$
$$V_{\text{bus}} = \beta_t \times t_{\text{bus}} + \beta_c \times c_{\text{bus}} + \varepsilon_{\text{bus}}$$
$$V_{\text{train}} = \beta_t \times t_{\text{train}} + \beta_c \times c_{\text{train}} + \varepsilon_{\text{train}}$$

However, there could be correlations among ε's across the **items**, for instance, customer who dislike public transportation will sample small(or even negative) utility for both bus and train. Additionally there could also be correlations among unobervables across the **individuals** as well.

We tackle this using either a **Nested Logit(NL) Model** or **Mixture Logit(ML) Model**. NL is used to capture the correlation in unobservables across the **items** and ML is used to capture the correlation in unobservables across **individuals**. Think of the above example, we can model it under an NL model as in Figure 1. The products in the same nest share some similarities and so that there probably are correlated between unobserved features.

4 Nested Logit Model

Although it can be generalized to d-level nested logit model, we focus on a 2-level one right now. Assume that products are partitioned into m nests. Each item belongs to **exactly one**
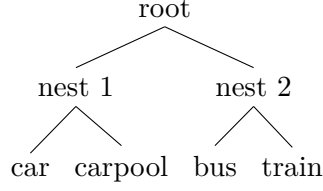
Figure 1: The Nested Logit Model for Example 3

nest. Each nest contains n items. Total number of products is mn.(We can also generalize it to the case where there are different number of products across the nests.) For each item j in nest ℓ, we model its random utility as following,

$$U_{\ell j} = W_\ell + Y_{\ell j} + \tilde{\varepsilon}_\ell + \tilde{\varepsilon}_{\ell j},$$

where

$$U_{\ell j} : \text{random utility of product } j \text{ in nest } \ell$$
$$W_\ell : \text{nest-specific utility for nest } \ell$$
$$Y_{\ell j} : \text{utility from item } j \text{ in nest } \ell$$
$$\tilde{\varepsilon}_\ell : \text{nest-specific unobserved utility for nest } \ell \text{(introduces correlation)}$$
$$\tilde{\varepsilon}_{\ell j} : \text{unobserved utility from item } j \text{ in nest } \ell$$

Then we have the following assumption for those error terms.

Assumptions:

1. $\tilde{\varepsilon}_\ell$, $\tilde{\varepsilon}_{\ell j}$ are independent across nests and products within nests.

2. $\tilde{\varepsilon}_{\ell j}$ are independent identically distributed Gumbel(0, μ_ℓ), $\mu_\ell > 0$.

3. $\tilde{\varepsilon}_\ell$ are distributed such that $\max_{j \in S_\ell} U_{j\ell}$ is Gumbel with scale $\lambda_\ell > 0$, for all the subsets of items S_ℓ in nest ℓ.

To better understand where the third assumption comes from, suppose we partition the four products as in Figure 2.

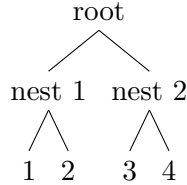
Figure 2: The Nested Logit Model for Example 4

Example 4. Under the NL model in Figure 2, the random utilities of four items can be modeled as

$$U_{11} = V_{11} + \tilde{\varepsilon}_1 + \tilde{\varepsilon}_{11}, \qquad\qquad U_{12} = V_{12} + \tilde{\varepsilon}_1 + \tilde{\varepsilon}_{12},$$
$$U_{23} = V_{23} + \tilde{\varepsilon}_2 + \tilde{\varepsilon}_{23}, \qquad\qquad U_{24} = V_{24} + \tilde{\varepsilon}_2 + \tilde{\varepsilon}_{24}.$$

In Assumption 3, we want

$$\begin{aligned}
\max\{U_{11}, U_{12}\} &= \max\{V_{11} + \tilde{\varepsilon}_1 + \tilde{\varepsilon}_{11}, V_{12} + \tilde{\varepsilon}_1 + \tilde{\varepsilon}_{12}\} \\
&= \tilde{\varepsilon}_1 + \max\{V_{11} + \tilde{\varepsilon}_{11}, V_{12} + \tilde{\varepsilon}_{12}\} \\
&= \tilde{\varepsilon}_1 + \mathrm{Gumbel}(\_\_, \mu_\ell) \overset{d}{\sim} \mathrm{Gumbel}(\_\_, \lambda_\ell).
\end{aligned}$$

More generally, denote $\mathbb{P}(j\ell) = \mathbb{P}(j \text{ from nest } \ell \text{ is chosen})$, $\mathbb{P}(\ell) = \mathbb{P}(\text{an item from nest } \ell \text{ is chosen})$, $\mathbb{P}(j|\ell) = \mathbb{P}(j \text{ is chosen}| \text{ item from nest } \ell \text{ is chosen})$. Thus,

$$\mathbb{P}(j\ell) = \mathbb{P}(\ell) \times \mathbb{P}(j|\ell),$$

where

$$\begin{aligned}
\mathbb{P}(\ell) &= \mathbb{P}(\max_{j \in N_\ell} U_{\ell j} \geq \max_{j \in N_{\ell'}} U_{\ell' j}, \forall \ell' \neq \ell) \\
&= \mathbb{P}(W_\ell + \tilde{\varepsilon}_\ell + \max_{j \in N_\ell}(Y_{\ell j} + \tilde{\varepsilon}_{\ell j}) \geq W_{\ell'} + \tilde{\varepsilon}_{\ell'} + \max_{j \in N_{\ell'}}(Y_{\ell' j} + \tilde{\varepsilon}_{\ell' j}), \forall \ell' \neq \ell)
\end{aligned}$$

Let $\max_{j \in N_\ell}(Y_{\ell j} + \tilde{\varepsilon}_{\ell j}) = \frac{1}{\mu_\ell} IV_\ell + \varepsilon_\ell$, where $\varepsilon_\ell \overset{d}{\sim} \mathrm{Gumbel}(0, \mu_\ell)$ and $IV_\ell = \log(\sum_{j \in N_\ell} e^{\mu_\ell Y_{\ell j}})$, where IV_ℓ is called the inclusive value/utility of nest ℓ, then

$$\mathbb{P}(\ell) = \mathbb{P}(W_\ell + \frac{1}{\mu_\ell} IV_\ell + \tilde{\varepsilon}_\ell + \varepsilon_\ell \geq W_{\ell'} + \frac{1}{\mu_{\ell'}} IV_{\ell'} + \tilde{\varepsilon}_{\ell'} + \varepsilon_{\ell'}, \forall \ell' \neq \ell)$$

We want $\tilde{\varepsilon}_\ell + \varepsilon_\ell$ to be independent and Gumbel of same scale to utilize the formula from MNL model and thus we assume that $\tilde{\varepsilon}_\ell + \varepsilon_\ell \overset{d}{\sim} \mathrm{Gumbel}(0, \lambda_\ell)$, which is exactly the third assumption we made for NL model. Then by formula for the MNL,

$$\mathbb{P}(\ell) = \frac{e^{\lambda_\ell(W_\ell + \frac{1}{\mu_\ell} IV_\ell)}}{\sum_{\ell'} e^{\lambda_{\ell'}(W_{\ell'} + \frac{1}{\mu_{\ell'}} IV_{\ell'})}}.$$